THE BATTLE OF NEIGHBOURHOODS

COURSERA – APPLIED DATA SCIENCE CAPSTONE

INTRODUCTION

- **Background**

  Toronto is the largest and the most populous city in Canada. It is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. The diverse population of Toronto reflects its current and historical role as an important destination for immigrants to Canada. Over 200 distinct ethnic origins are represented among its inhabitants and while the majority of population speak English as their primary language, over 160 languages are spoken in the city.

  Given the population and diversity of Toronto, the market is mostly saturated and leads to a competitive environment for business owners. Starting a new business will be more successful if extensive analysis is done in a structured manner and insights will help to choose the right business model and structure.

- **Problem Statement**

  In a city as diverse and developed like Toronto, opening a yoga studio is an exciting but also an exhausting task. Once decided, the next step should be finding an ideal location for the studio. Multiple factors are needed to be analysed for the best location which include but not limited to:

  - Style of operation
  - Demographics
  - Foot traffic
  - Competition

  It's crucial to find the best location for a yoga studio and if successful same recommendations can be used to expand into other neighbourhoods of Toronto.

- **Target Audience**

  The research is going to be interested by anyone who is planning to open a yoga studio in Toronto. Goal is to identify and recommend which neighbourhood(s) should be the targeted to open a yoga studio.

DATA

- **Toronto Neighbourhoods**

  A list boroughs and neighbourhoods of Toronto is essential in order to analyse the neighbourhoods. This dataset will be used to capture the coordinates of specific neighbourhoods.

  Dataset is available on Toronto Open Data website at https://open.toronto.ca/dataset/neighbourhoods/. CSV file can be found at https://ckan0.cf.opendata.inter.prod-toronto.ca/download_resource/a083c865-6d60-4d1d-b6c6-b0c8a85f9c15?format=csv&projection=4326.

- **Toronto Demographics**

  Capturing neighbourhoods and their demographics are critical. Multiple factors will be included in the analysis, such as population, density, average income. Demographics dataset will be enriched with Toronto neighbourhoods for further insights.

  Dataset is available on Wikipedia at https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods.

- **List of Places in Toronto**

  Capturing neighbourhoods and their demographics are critical. Multiple factors will be included in the analysis, such as population, density, average income. Demographics dataset will be enriched with Toronto neighbourhoods for further insights.

  Dataset is available on Wikipedia at https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods.

DATA CAPTURE AND CLEANING

Toronto neighbourhood data is scraped and downloaded from multiple sources like Wikipedia, Toronto Open Data, and Foursquare. Multiple features are dropped from Toronto Open Data dataset since only latitude and longitude values will be required. Demographics dataset is scraped from Wikipedia and contains 12 features. Only neighbourhood, borough, and population are kept and others are ignored for the purpose of this report.

Combining demographics and coordinates dataset requires cleaning as the neighbourhood names do not have one-to-one matches. Whenever a neighbourhood name is found is coordinates dataset (Figure 1), those coordinates are used to calculate the mean and the result is assigned as neighbourhood coordinate (Figure 2). represents a sample of coordinate calculation.

Toronto Neighbourhood Coordinates, Toronto Open Data

Neighbourhood	Lat	Lng
Agincourt North	43.8054405769	-79.266712166
Agincourt South-Malvern West	43.788657551099995	-79.2656117966

Figure 1 - Toronto Neighbourhood Coordinates Data Sample

Aggregated Neighbourhood Coordinates

Neighbourhood	Lat	Lng
Agincourt	43.79704906399999	-79.2661619813

Figure 2 - Aggregated Neighbourhood Coordinates

DATA CAPTURE AND CLEANING

Neighbourhoods that have coordinates available are kept and the rest is ignored since coordinates are mandatory in order to utilize Foursquare APIs.

Using Foursquare Place API, two separate sets of data are gathered; yoga studios & gyms (Figure 3), and nearby stores, residentials, shops, and offices (Figure 4). Although the data fetched from Foursquare contains many features, only name, coordinates, and category are used.

Studios in Toronto, Foursquare Places API

Neighbourhood	Venue	Venue Lat	Venue Lng	Venue Category	Venue Top Category
Amesbury	Nama Saltra	43.70148086547852	-79.4766845703125	Yoga Studio	Yoga Studio
Playter Estates	Riverdale Pilates	43.67792874700868	-79.35036471041796	Pilates Studio	Pilates Studio
Casa Loma	Spynga	43.68140509016735	-79.41747561982021	Yoga Studio	Gym / Fitness Center

Figure 3 - Yoga Studios & Gyms in Toronto

Places in Toronto, Foursquare Places API

Neighbourhood	Venue	Venue Lat	Venue Lng	Venue Category	Venue Top Category
Agincourt	Twilight	43.791999483291015	-79.25858447143591	Lounge	Nightlife Spot
Kingsview Village	Jones DesLauriers Insurance	43.705344333928736	-79.55166399478912	Office	Professional & Other Places
Bayview Village	Taro's Fish	43.769961643415144	-79.37465906198047	Fish Market	Shop & Service

Figure 4 - Nearby Places in Toronto

DATA CAPTURE AND CLEANING

In order to prepare the Foursquare data for consumption, an average value of each categories per neighbourhood is calculated and plotted. The averages represent the density and distribution of different categories in the neighbourhoods being analysed. Since the datasets are divided in 2 set; studios and places, they are processed individually (Figure 5, Figure 6) and combined (Figure 7).

Studios Distribution Per Neighbourhood

Neighbourhood	Gym / Fitness Center	Gymnastics Gym	Pilates Studio	Yoga Studio
Agincourt	1.0	0.0	0.0	0.0
Alderwood	1.0	0.0	0.0	0.0
Amesbury	0.8333333333333334	0.0	0.0	0.16666666666666666

Figure 5 - Studio Distribution Per Neighbourhood

Places Distribution Per Neighbourhood

Neighbourhood	Nightlife Spot	Professional & Other Places	Residence	Shop & Service
Agincourt	0.07692307692307693	0.19230769230769232	0.0	0.7307692307692307
Alderwood	0.11428571428571428	0.45714285714285713	0.05714285714285714	0.37142857142857144
Amesbury	0.0	0.36363636363636365	0.09090909090909091	0.5454545454545454

Figure 6 – Places Distribution Per Neighbourhood

Combined Distribution Per Neighbourhood

Neighbourhood	Gym / Fitness Center	Gymnastics Gym	Pilates Studio	Yoga Studio	Nightlife Spot	Professional & Other Places	Residence	Shop & Service
Agincourt	1.0	0.0	0.0	0.0	0.07692307692307693	0.19230769230769232	0.0	0.7307692307692307
Alderwood	1.0	0.0	0.0	0.0	0.11428571428571428	0.45714285714285713	0.05714285714285714	0.37142857142857144
Amesbury	0.8333333333333334	0.0	0.0	0.16666666666666666	0.0	0.36363636363636365	0.09090909090909091	0.5454545454545454

Figure 7 - Combined Distribution Per Neighbourhood

DATA EXPLORATION

Visualization of Toronto and its neighbourhoods shown on a map (Figure 8). The distribution and count of neighbourhoods are pleasant and each neighbourhood is represented with blue colour.

Visualization of Studios (Figure 9) and Places (Figure 10) along with the neighbourhoods. Blue colour represents the neighbourhoods, red colour represents the places, yellow colour represents the yoga studios, and green colour represents other studios such as gyms, fitness centres. Most yoga studios are concentrated around downtown Toronto as seen on the map.

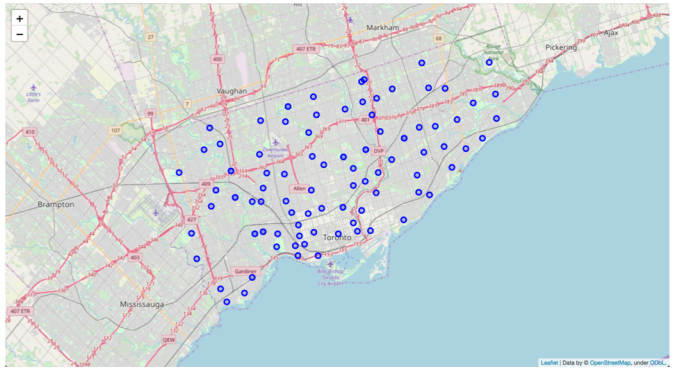

Figure 8 - Toronto Neighbourhoods Map

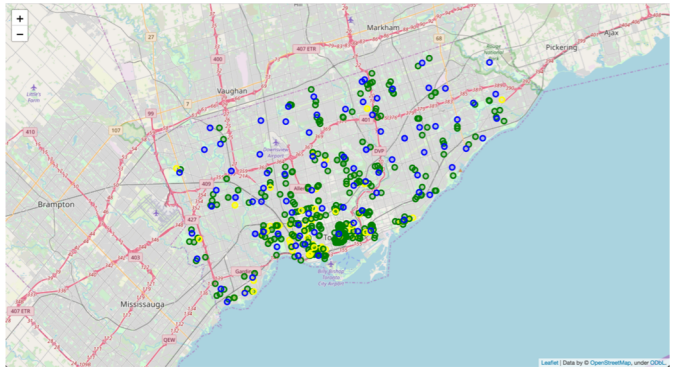

Figure 9 - Toronto Studios Distribution Map

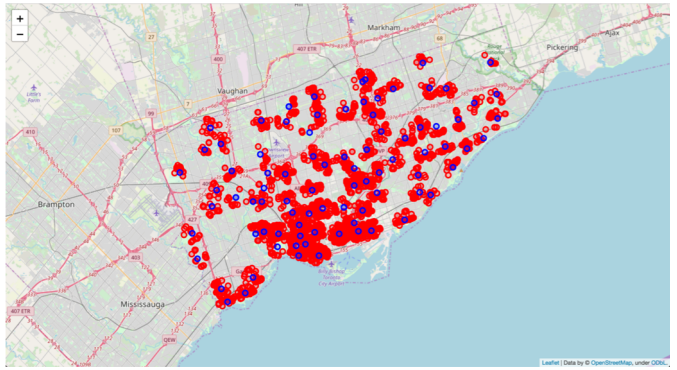

Figure 10 - Toronto Places Distribution Map

DATA EXPLORATION

Once key datasets are identified and visualized, the neighbourhoods are being clustered using K-Means clustering algorithm. Selection of optimum k-values are decided by using The Elbow Method (Figure 11). The best k-value for Studios picked as 3, for Places picked as 5, and for Combined dataset picked as 5.

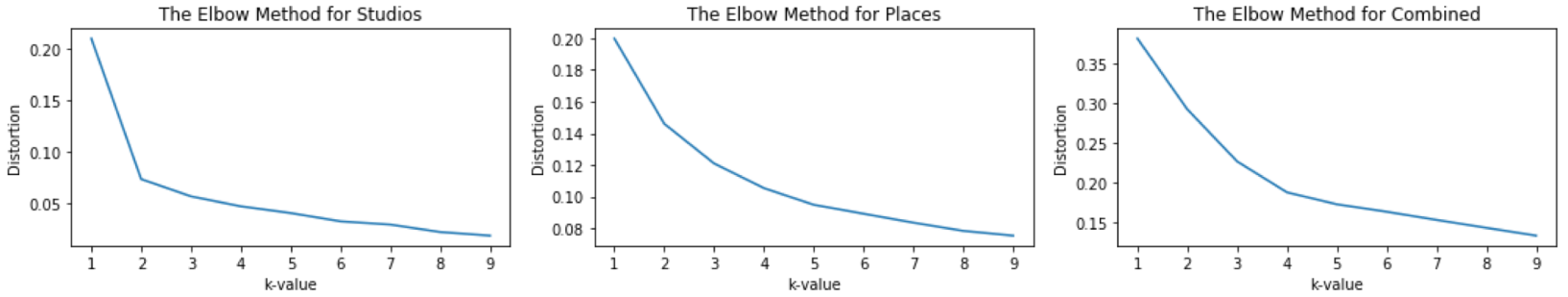

Figure 11 - The Elbow Method

DATA EXPLORATION

Clustering neighbourhoods based on Studios provides 3 distinct clusters. A sample result is provided (Figure 12) and visualized on a map (Figure 13) where red colour represents Cluster 0, purple colour represents Cluster 1, and teal colour represents Cluster 2.

Based on the density of each category in the clusters (Figure 14), Cluster 1 has almost no yoga studios. Although both Cluster 0 and Cluster 2 larger values, yoga studios are mostly concentrated in Cluster 2. The same information can also be seen on the map (Figure 15) where red colour represents Cluster 0, purple colour represents Cluster 1, teal colour represents Cluster 2, yellow colour represents yoga studios, and green colour represents other studios such as gyms, fitness centres.

Neighbourhood Clusters based on Studios

Neighbourhood	Cluster Labels	Gym / Fitness Center	Gymnastics Gym	Pilates Studio	Yoga Studio
Agincourt	1	1.0	0.0	0.0	0.0
Bedford Park	0	0.7142857142857143	0.14285714285714285	0.0	0.14285714285714285
Casa Loma	2	0.45454545454545453	0.0	0.18181818181818182	0.36363636363636365

Figure 12 - Neighbourhood Clusters based on Studios

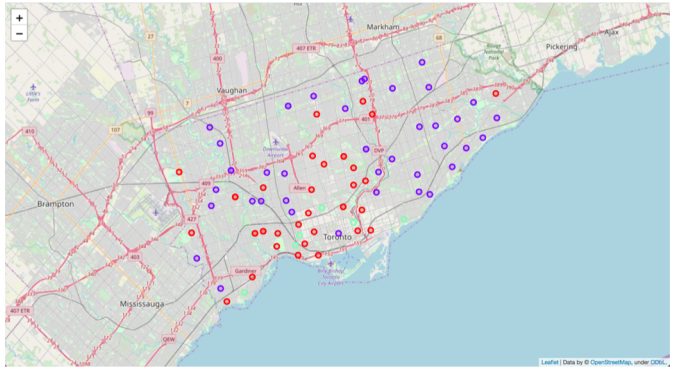

Figure 13 - Neighbourhood Clusters Map based on Studios

Category Density Per Cluster based on Studios

Cluster Labels	Gym / Fitness Center	Gymnastics Gym	Pilates Studio	Yoga Studio
0.0	0.7275688560410783	0.013227513227513227	0.028487921543477093	0.23071570918793138
1.0	0.9976009596161535	0.0003998400639744102 4	0.0003998400639744102 4	0.001599360255897641
2.0	0.47746697746697747	0.0	0.09654234654234654	0.42599067599067597

Figure 14 - Category Density Per Cluster based on Studios

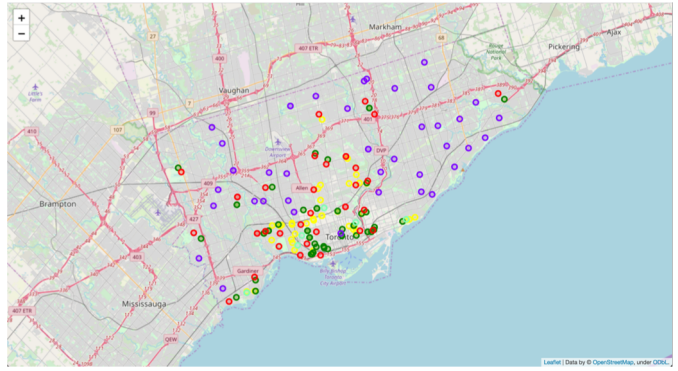

Figure 15 - Neighbourhood Clusters Map based on Studios, along with Studios

DATA EXPLORATION

Clustering neighbourhoods based on Places provides 5 distinct clusters. A sample result is provided (Figure 16) and visualized on a map (Figure 17) where red colour represents Cluster 0, purple colour represents Cluster 1, blue colour represents Cluster 2, teal colour represents Cluster 3, and orange colour represents Cluster 4.

Based on the density of each category in the clusters (Figure 18), Cluster 0 and Cluster 4 have the lowest density of nightlife spots. All of the clusters have very low density of residences and mainly populated with shop & service and professional & other places. The same information can also be seen on the map (Figure 19) where red colour represents Cluster 0, purple colour represents Cluster 1, blue colour represents Cluster 2, teal colour represents Cluster 3, orange colour represents Cluster 4, and green colour represents places.

Neighbourhood Clusters based on Places

Neighbourhood	Cluster Labels	Nightlife Spot	Professional & Other Places	Residence	Shop & Service
Agincourt	2	0.07692307692307693	0.1923076923076923 2	0.0	0.7307692307692307
Bayview Village	4	0.018518518518518517	0.5370370370370371	0.05555555555555555	0.3888888888888889
Cabbagetown	0	0.08620689655172414	0.3017241379310345	0.15517241379310345	0.45689655172413796

Figure 16 - Neighbourhood Clusters based on Places

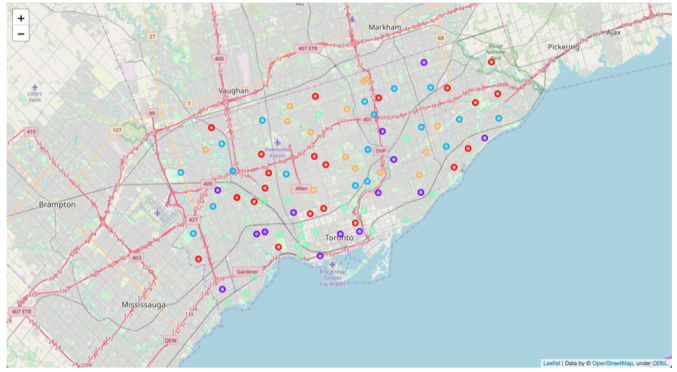

Figure 17 - Neighbourhood Clusters Map based on Places

Category Density Per Cluster based on Places

Cluster Labels	Nightlife Spot	Professional & Other Places	Residence	Shop & Service
0.0	0.045684764156646566	0.3789961174905551	0.03983132589390714	0.5354877924588911
1.0	0.13726015729509009	0.415063958809104	0.08757138326489453	0.36010450063091143
2.0	0.06299303438788581	0.19812035309901416	0.02645234090537431	0.7124342716077257
3.0	0.1831780195212287	0.28891086778459757	0.019238148674367468	0.5086729640198063
4.0	0.04107288215795428	0.5961000880405799	0.05277408754820014	0.3055496210466458

Figure 18 - Category Density Per Cluster based on Places

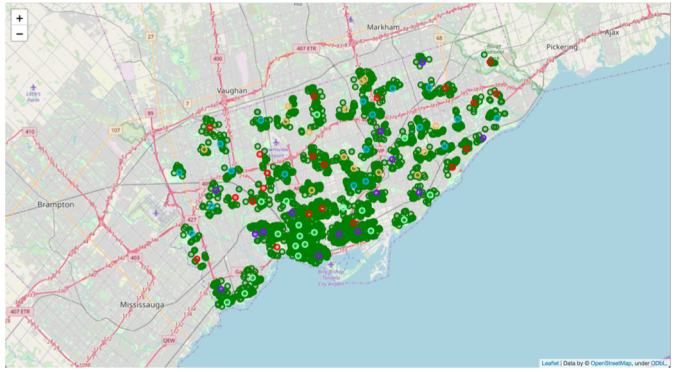

Figure 19 - Neighbourhood Clusters Map based on Places, along with Places

DATA EXPLORATION

Clustering neighbourhoods based on Studios and Places combined provides 5 distinct clusters. A sample result is provided (Figure 20) and visualized on a map (Figure 21) where red colour represents Cluster 0, purple colour represents Cluster 1, blue colour represents Cluster 2, teal colour represents Cluster 3, and orange colour represents Cluster 4.

Based on the density of each category in the clusters (Figure 22), in all the clusters density of pilates studios and gymnastics gyms are very low and sometimes non-existent. Nightlife spots, gyms, and yoga studios have the highest density in Cluster 2 and Cluster 3. Professional & other places and residences are higher in density in Cluster 0 and Cluster 4. The same information can also be seen on the map (Figure 23) where red colour represents Cluster 0, purple colour represents Cluster 1, blue colour represents Cluster 2, teal colour represents Cluster 3, and orange colour represents Cluster 4, yellow colour represents yoga studios, and green colour represents places and other studios such as gyms, fitness centres.

Figure 20 - Neighbourhood Clusters based on both Studios and Places

Figure 22 - Category Density Per Cluster based on Studios and Places

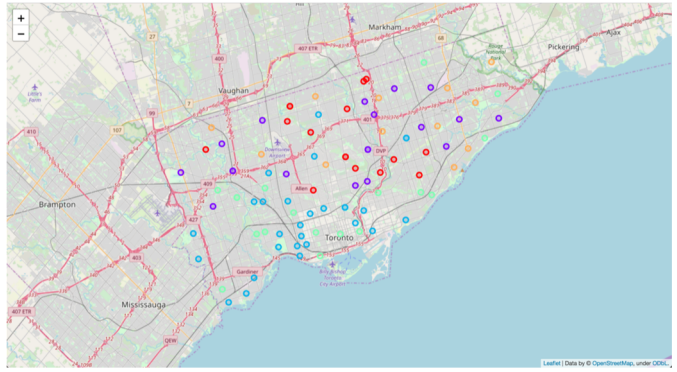

Figure 21 - Neighbourhood Clusters Map based on both Studios and Places

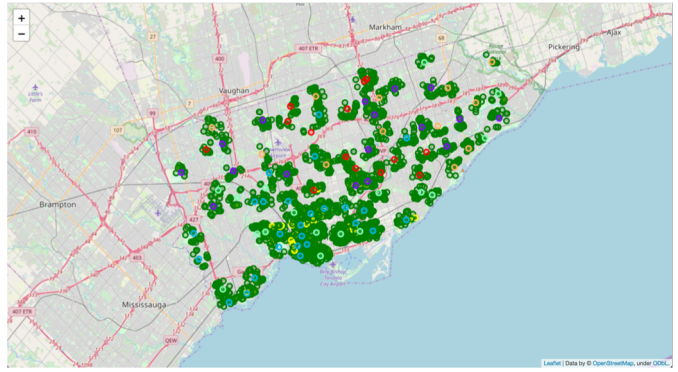

Figure 23 - Neighbourhood Clusters Map based on Studios and Places, along with Studios and Places

RESULTS

By visualizing the combined clusters and the density of yoga studios overlapping with those clusters (Figure 24), Cluster 2 and Cluster 3 have higher concentration of yoga studios compared to other clusters where red colour represents Cluster 0, purple colour represents Cluster 1, blue colour represents Cluster 2, teal colour represents Cluster 3, orange colour represents Cluster 4, and yellow colour represents yoga studios. Hence, assuming the business is able to develop in those neighbourhood, opening a yoga studio in one of these clusters will have a higher success rate.

Further analysis of Cluster 2 and Cluster 3 visualized on a map (Figure 25), where blue colour represents Cluster 2, teal colour represents Cluster 3, and yellow colour represents yoga studios. In both clusters most of the yoga studios are concentrated in downtown Toronto and the market is more saturated. However, there are some untapped neighbourhoods available outside of downtown which makes those neighbourhoods a good candidate for a new yoga studio.

As seen on the map, most of the yoga studios are nearby Cluster 2 neighbourhoods, and red marked two neighbourhoods (Figure 26) is a good candidate for a new yoga studio since they share a lot of similar features where yoga studios are concentrated and since two neighbourhoods are close to each other which attract customers from both neighbourhoods.

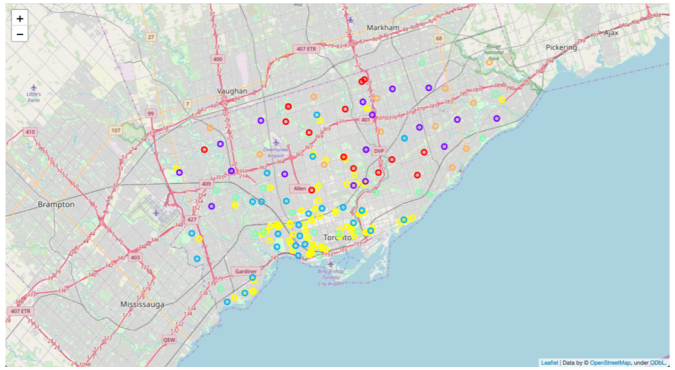

Figure 24 - Neighbourhood Clusters Map based on Studios and Places, along with Yoga Studios

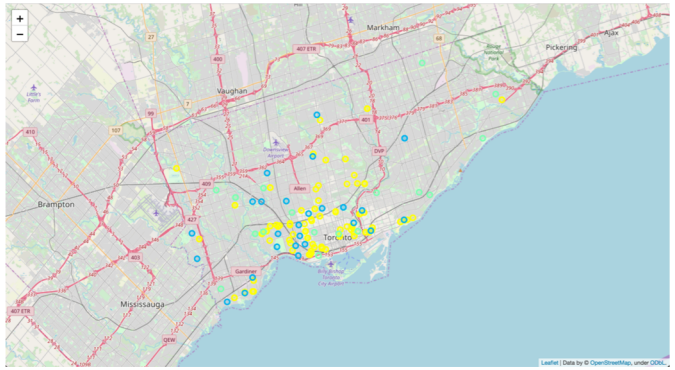

Figure 25 - Map of Cluster 2 and Cluster 3, along with Yoga Studios

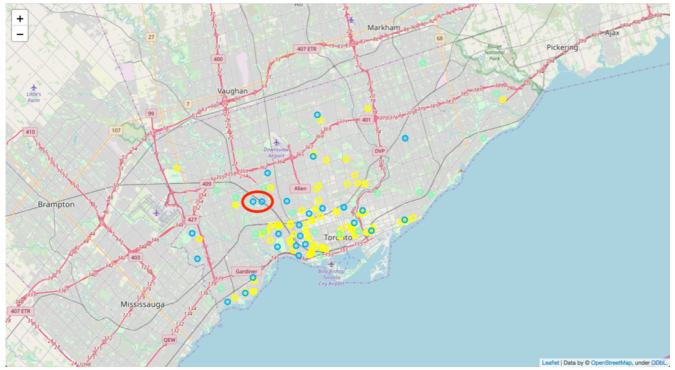

Figure 26 - Best Neighbourhood Candidates

DISCUSSION AND CONCLUSION

- **Discussion**

  Toronto is a large city with 140 neighbourhoods. Due to the inconsistency of coordinate dataset and demographics dataset, only a subset of those neighbourhoods captured. Improving the consistency may result in different clusters and hence different results.

  During the feature selection, some features are not included such as population of the neighbourhoods or the average income. Including more features that improves the clustering algorithm is a good addition to improve the overall recommendation. It's also important not to introduce a lot of features that are overlapping since they tend to skew the clusters.

- **Conclusion**

  In this report, neighbourhoods of Toronto are analysed with the goal of identifying best fitting neighbourhoods that are favourable opening a yoga studio. Clustering models are used in order to find similar neighbourhoods and then compared to the existing yoga studios overlapping with those clusters. The models and methodology can be useful to identify the best neighbourhood for a yoga studio but also can be enhanced or modified to identify other categories based on the needs.